IMPOSTURE USING SYNTHETIC SPEECH AGAINST SPEAKER VERIFICATION BASED ON SPECTRUM AND PITCH

Takashi Masuko†, Keiichi Tokuda††, and Takao Kobayashi†

†Interdisciplinary Graduate School of Science and Engineering,
Tokyo Institute of Technology, Yokohama, 226-8502 JAPAN
††Department of Computer Science, Nagoya Institute of Technology, Nagoya, 466-8555 JAPAN

ABSTRACT

This paper describes security of speaker verification systems against imposture using synthetic speech. We propose a text-prompted speaker verification technique which utilizes pitch information in addition to spectral information, and investigate whether synthetic speech is rejected. Experimental results show that pitch information is not necessarily useful for rejection of synthetic speech, and it is required to develop techniques to discriminate synthetic speech from natural speech.

1. INTRODUCTION

For speaker verification systems, security against imposture is one of the most important problems, and many approaches to reducing false acceptance rates for impostors as well as false rejection rates for clients have been investigated. For example, text-prompted speaker verification techniques [1] are robust to the impostor with playing back recorded voice of a registered speaker. However, imposture using synthetic speech has barely been taken into account due to the facts that quality of the synthetic speech was not enough, and that it was difficult to synthesize speech with arbitrary voice characteristics.

Meanwhile, recent advances in speech synthesis make it possible to synthesize speech of good quality. We have also proposed an HMM-based speech synthesis system [2],[3] which can synthesize smooth and natural sounding speech. Moreover, we have shown that we can change voice characteristics of synthetic speech to resemble target speaker's voice characteristics by applying speaker adaptation techniques using a small amount of adaptation data [4]. From these points of view, we presented preliminary experimental results on imposture against speaker verification systems using the HMM-based speech synthesis system [5]. In [5], we used an HMM-based text-prompted speaker verification system as a reference system, and showed that false acceptance rates for synthetic speech reached over 70% by training the synthesis system using only one sentence from each customer, while a false acceptance rate for human impostors was 0%. However, the experimental conditions were not necessarily realistic. For example, synthetic speech was generated using white noise excitation without pitch information.

In this paper, to reject speech synthesized without pitch, we utilize pitch information for speaker verification. Pitch patterns and spectral parameters can be modeled simultaneously by multi-space probability distribution HMM (MSD-HMM) [6]. Based on the MSD-HMM, we construct a text-prompted speaker verification system. To model variations of pitch patterns accurately, contextual factors, such as phoneme identity factors or stress-related factors, are taken into account. Then, we investigate whether the speaker verification system can reject synthetic speech from the HMM-based speech synthesis system which is trained using speech data from customers of the speaker verification system.

This paper is organized as follows. In Section 2, a text-prompted speaker verification system based on MSD-HMM is explained briefly. The experimental conditions and results are shown in Section 3 and 4, and the conclusion is given in Section 5.

2. TEXT-PROMPTED SPEAKER VERIFICATION BASED ON SPECTRUM AND PITCH

The observation sequence of pitch pattern consists of one-dimensional continuous values and discrete symbol which represents "unvoiced". The main problem of pitch pattern modeling is how to model these observations which have quite different properties. Several approaches have been proposed to utilize pitch information for (text-independent) speaker recognition. For example, in [7] and [8], the score for pitch part is calculated from pitch values within voiced regions and combined with the score for spectral part, and in [9] and [10], two speaker models, which corresponds to voiced and unvoiced regions respectively, are constructed for each speaker. In this paper, we adopt an alternative ap-

This work was partially supported by the Ministry of Education, Science and Culture of Japan, Grant-in-Aid for Encouragement of Young Scientists, 09750399, 1997.

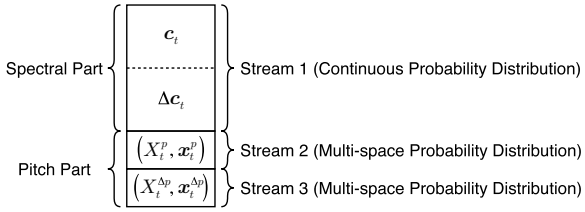


Figure 1: Observation vector.

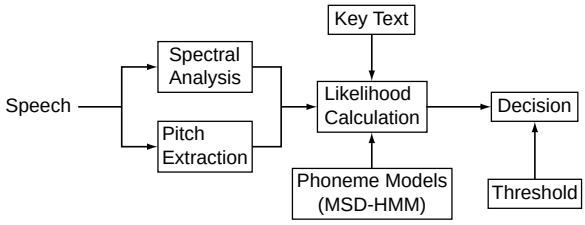


Figure 2: Blockdiagram of a speaker verification system based on MSD-HMM.

proach in which pitch observation sequences are modeled by the multi-space probability distribution HMM (MSD-HMM).

MSD-HMM can model pitch observations strictly without any heuristic assumptions, and a reestimation algorithm for MSD-HMM is derived in [6]. Therefore, we can apply MSD-HMM to traditional HMM-based text-prompted speaker verification directly. MSD-HMM can model variations of pitch observations accurately by taking account of linguistic contexts as well as phonetic contexts, and a decision-tree based context clustering technique [11] is extended for MSD-HMM [12] which enables to balance model complexity against data availability.

2.1. Modeling Pitch Observations Using MSD-HMM

We assume that pitch pattern is a sequence of outputs from a one-dimensional space Ω_1 and a zero-dimensional space Ω_2 which correspond to voiced and unvoiced regions, respectively. Each space Ω_g has its probability w_g, i.e., probability for voiced observation w_1 and for unvoiced observation w_2, where $\sum_{g=1}^{2} w_g = 1$. The space Ω_1 has a one-dimensional probability density function $\mathcal{N}_1(\boldsymbol{x})$ where $\int_{\Omega_1} \mathcal{N}_1(\boldsymbol{x}) d\boldsymbol{x} = 1$, and Ω_2 has only one sample point. A pitch observation $\boldsymbol{o}$ consists of a continuous random variable $\boldsymbol{x}$ and a set of spaces indices X, that is,

$$\boldsymbol{o} = (X, \boldsymbol{x}), \tag{1}$$

where $X = \{1\}$ for voiced region and $X = \{2\}$ for unvoiced region. The observation probability of $\boldsymbol{o}$ is defined by

$$b(\boldsymbol{o}) \quad = \quad \sum_{g \in X} w_g \mathcal{N}_g(\boldsymbol{x}). \tag{2}$$

It is noted that, although $N_2(\boldsymbol{x})$ does not exist for Ω_2, we define as $N_2(\boldsymbol{x}) \equiv 1$ for simplicity of notation.

Here we consider an HMM whose output probability in each state is given by equation (2). We call this type of HMM MSD-HMM. Using MSD-HMM, We can model voiced and unvoiced observations of pitch in a unified model without any heuristic assumption. Moreover, we can model spectrum and pitch simultaneously using multi-stream MSD-HMM, in which spectral part is modeled by continuous probability distribution, and pitch part is modeled by MSD (Figure 1).

2.2. Text-prompted Speaker Verification Based on MSD-HMM

A blockdiagram of a text-prompted speaker verification system based on MSD-HMM is shown in Figure 2. In the training stage, a set of phoneme models is trained for each customer. To model variations of pitch patterns accurately, phonetic and linguistic contexts are taken into account, and a decision-tree based context clustering technique is applied to the context dependent models. A set of speaker and context independent phoneme models is also trained using all the customers' training data.

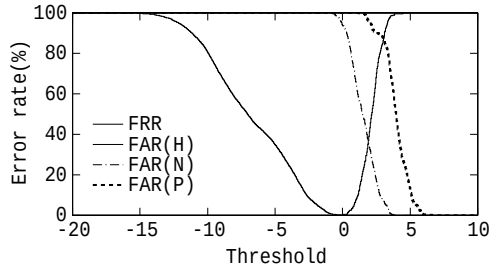
In the verification stage, mel-cepstral coefficients and a logarithm of the fundamental frequency are extracted, and their delta parameters are calculated. Then, normalized log-likelihood of input parameter sequence $\boldsymbol{O}$ for the claimant speaker s is calculated as follows,

$$L_s(\boldsymbol{O}) = \frac{1}{T} \left\{ \log P(\boldsymbol{O}|w, \lambda_s) - \max_{v \in W} \log P(\boldsymbol{O}|v, \lambda_{SI}) \right\}, \tag{3}$$

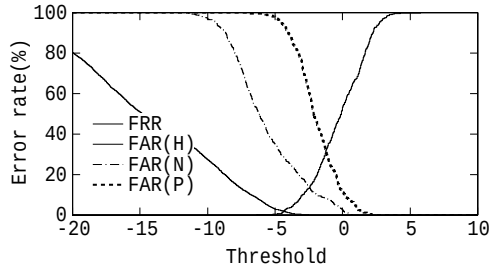
where T is the total number of frames of input speech, w is the label sequence corresponding to the key text presented to the speaker, λ_s is a set of phoneme models of the claimant speaker, W is a set of possible label sequence, λ_{SI} is a set of speaker independent phoneme models, respectively.

3. EXPERIMENTAL CONDITIONS

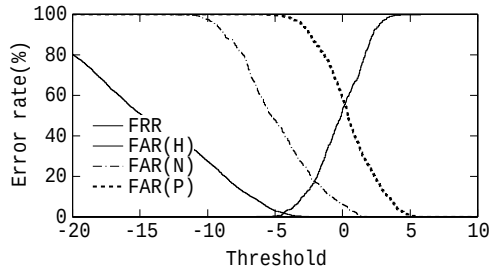
We used phonetically balanced Japanese sentences from ATR Japanese speech database. The database consists of sentence data uttered by ten speakers (six male speakers and four female speakers). All the speakers were used as customers. Customers except for the claimant speaker were also used as human impostors. Speech signals were sampled at 16kHz, and labeled into context dependent phoneme labels based on phoneme labels and linguistic information included in the database. We used 42 phonemes including silence and pause. The details of contextual factors are shown in [3]. Both the speech synthesis system and the speaker verification system used the same label sequences. We used 3-state left-to-right models with single diagonal Gaussian output distributions (for spectral part) for both the speech synthesis and the speaker verification systems.



(a) Speaker verification: CI, Speech Synthesis: CI



(b) Speaker verification: CD, Speech Synthesis: CI



(c) Speaker verification: CD, Speech Synthesis: CD

FAR(H): FAR for human impostor
FAR(N): FAR for speech synthesized without pitch
FAR(P): FAR for speech synthesized with pitch

Figure 3: False rejection and acceptance rates as functions of the values of the decision threshold.

For the speaker verification system, 100 sentences were used to train a set of phoneme models of each speaker. Speech signals were windowed by a 25 ms Blackman window with a 5 ms shift, and the cepstral coefficients were calculated by 20-th order LPC analysis. Pitch values were obtained using ESPS *get_f0* program [13]. The feature vector consisted of 20 cepstral coefficients including 0-th coefficient, a logarithm of the fundamental frequency, and their delta parameters. Delta pitch parameters were calculated only within voiced regions, and the frames, where delta pitch parameters were not computable because of the boundaries of voiced and unvoiced regions, were treated as unvoiced. Both context independent (CI) and context dependent (CD) models were trained for each speaker. Speaker independent CI models were also trained. In the verification stage, likelihood in equation (3) is calculated on the Viterbi path.

The speech synthesis system were trained using 50 sentences for each customer. These sentences did not overlap to the training data of the speaker verification system. Speech signals were windowed by a 25 ms Blackman window with a 5 ms shift, and the mel-cepstral coefficients were calculated by 24-th order mel-cepstral analysis [14]. Pitch values were obtained using *get_f0* program. The feature vector consisted of 24 mel-cepstral coefficients including 0-th coefficient, logarithm of fundamental frequency, and their deltas. As well as the speaker verification system, both CI and CD models were trained for each speaker. As the test data, 50 sentences were synthesized from both CI and CD models, and both with and without pitch. These sentences did not overlap to the training data.

Table 1: Equal error rates (speech synthesis system were trained using 50 sentences).

			speaker verification			
			w/o pitch		with pitch	
			CI	CD	CI	CD
human impostor			0.2	0.9	0.1	1.8
synthetic speech	w/o pitch	CI	82.0	37.8	36.0	13.4
		CD	71.4	53.8	29.6	17.4
	with pitch	CI	84.8	40.4	87.0	32.0
		CD	77.6	54.6	76.4	55.6

4. RESULTS

Figure 3 shows false rejection rates (FRRs) for customers and false acceptance rates (FARs) for human impostors and synthetic speech trained using 50 sentences as functions of the values of the decision threshold, and Table 1 shows the equal error rates (EERs). Figure 3 (a) shows the results where both the speaker verification and the speech synthesis systems used CI models, (b) shows the results where the speaker verification system used CD models, and (c) shows the results where the speech synthesis systems also used CD models.

EERs for synthetic speech generated using white noise excitation (without pitch) were reduced significantly by utilizing pitch information for speaker verification, however, EERs were considerably higher than human impostors, and reached over 10%. This could be attributed to higher likelihood for spectral part, i.e., the total likelihood is still high because of higher likelihood for spectral part, even though likelihood for pitch part is low. Furthermore, from the fact that EERs for speech synthesized with pitch reached over 30%, pitch information is hardly useful for speaker verification to reject synthetic speech with pitch. One of the reasons could be that the same pitch modeling technique as the speech synthesis system was used in the speaker verification system.

From the fact that EERs were reduced by using CD models in the speaker verification system, it could be considered that taking account of linguistic context is useful for speaker verification as well as for speech recognition. How-

Table 2: Equal error rates for synthetic speech with less training data.

synthetic speech		speaker verification			
No. of training sentences	Excitation	w/o pitch		with pitch	
		CI	CD	CI	CD
10 sentences	w/o pitch	49.0	32.6	14.8	11.4
	with pitch	64.0	33.6	52.2	29.8
5 sentences	w/o pitch	30.6	22.6	11.8	11.0
	with pitch	36.8	22.4	29.8	20.0
3 sentences	w/o pitch	24.2	15.8	5.8	8.8
	with pitch	24.6	15.2	19.0	11.8

ever, from Figure 3 (c), it can be seen that distributions of likelihood for customers and synthetic speech with pitch is overlapping, and EERs for synthetic speech from CD models with pitch are reached over 50%. From these results, using CD models for speaker verification is insufficient to reject synthetic speech.

Table 2 shows the EERs for synthetic speech with less training data. In these cases, we used CD models for speech synthesis. From Table 2, it can be seen that EERs for synthetic speech were more than 20% even though the HMM-based speech synthesis system was trained using only five sentences.

These experiments might be slightly biased against the speaker verification system. For example, both the speaker verification system and the speech synthesis system adopted the same approach to modeling pitch patterns, the distribution of spectral parameter was modeled by only one Gaussian distribution in each state, and many experimental conditions were identical between the speaker verification system and the speech synthesis system. However, the EERs for synthetic speech were considerably higher than those for impostors' speech. These results suggest that adjustment of decision threshold will not be able to reject synthetic speech effectively without significant increase of FRRs for customers, and some techniques to discriminate synthetic speech from natural speech are required.

5. CONCLUSION

In this paper, we have proposed a speaker verification technique which utilizes pitch information, and have investigated whether the speaker verification system can reject synthetic speech. The experiments might be slightly biased against the speaker verification system. However, from the facts that the false acceptance rates for synthetic speech with pitch reached over 20% by training the speech synthesis system using only five sentences for each speaker, current security of HMM-based speaker verification systems against synthetic speech is inadequate even though these disadvantages are taken into account.

To put speaker verification systems into practice, it is required to develop techniques to discriminate synthetic speech from natural speech. Investigation in another conditions such as speaker verification systems with different frameworks, and investigation on stand-alone speaker verification systems are also our future works.

6. REFERENCES

1. T. Matsui and S. Furui, "Speaker adaptation of tied-mixture-based phoneme models for text-prompted speaker recognition," Proc. ICASSP-94, pp.125–128, Apr. 1994.
2. T. Masuko, K. Tokuda, T. Kobayashi and S. Imai, "Speech synthesis using HMMs with dynamic features," Proc. ICASSP-96, pp.389–392, May 1996.
3. T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," Proc. EUROSPEECH'99, pp.2347–2350, Sep. 1999.
4. M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Speaker adaptation for HMM-based speech synthesis system using MLLR," Proc. The Third ESCA/COSCOSDA International Workshop on Speech Synthesis, pp.273–276, Nov. 1998.
5. T. Masuko, T. Hitotsumatsu, K. Tokuda, and T. Kobayashi, "On the security of HMM-based speaker verification systems against imposture using synthetic speech," Proc. EUROSPEECH-99, pp.1223–1226, Sep. 1999.
6. K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," Proc. ICASSP'99, pp.229–232, Mar. 1999.
7. M. J. Carey, E. S. Parris, H. L. Thomas, and S. Bennett, "Robust prosodic features for speaker identification," Proc. ICSLP-96, pp.1800–1803, Oct. 1996
8. M. K. Sönmez, L. Heck, M. Weintraub, and E. Shriberg, "A lognormal tied mixture model of pitch for prosody-based speaker recognition," Proc. EUROSPEECH-97, pp.1391–1394, Sep. 1997.
9. T. Matsui and S. Furui, "Text-independent speaker recognition using vocal tract and pitch information," Proc. ICSLP-90, pp.137–140, Nov. 1990.
10. K. P. Markov and S. Nakagawa, "Text-independent speaker recognition using multiple information sources," Proc. ICSLP-98, pp.173–176, Dec. 1998.
11. J. J. Odell, "The use of context in large vocabulary speech recognition," PhD dissertation, Cambridge University, Mar. 1995.
12. T. Masuko, K. Tokuda, N. Miyazaki, and T. Kobayashi, "Pitch Pattern Generation Using Multi-space Probability Distribution HMM," Trans. IEICE, J83-D-II, 7, pp.1600–1609, July 2000 (in Japanese).
13. Entropic Research Laboratory, Inc., *ESPS Programs Version 5.0*, 1993.
14. T. Fukada, K. Tokuda, T. Kobayashi and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," Proc. ICASSP-92, pp.137–140, Mar. 1992.